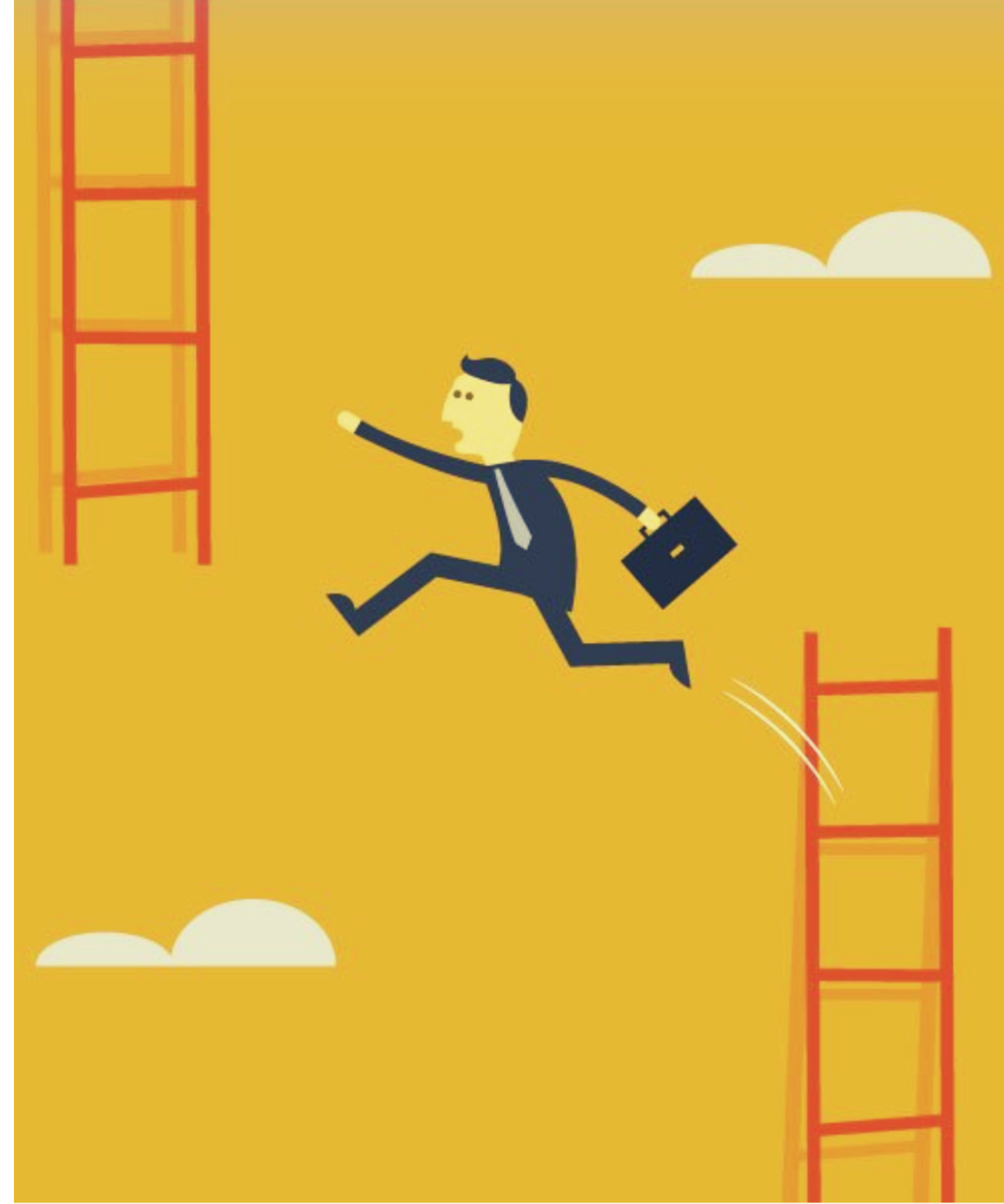


# Data Scientist Job Change Prediction

By Raghad Althunayan & Shatha Almoteb



# Table of Contents

1. Introduction

2. Workflow

3. Data & Design

4. Models

5. Model Deployment

6. Tools

7. Conclusion

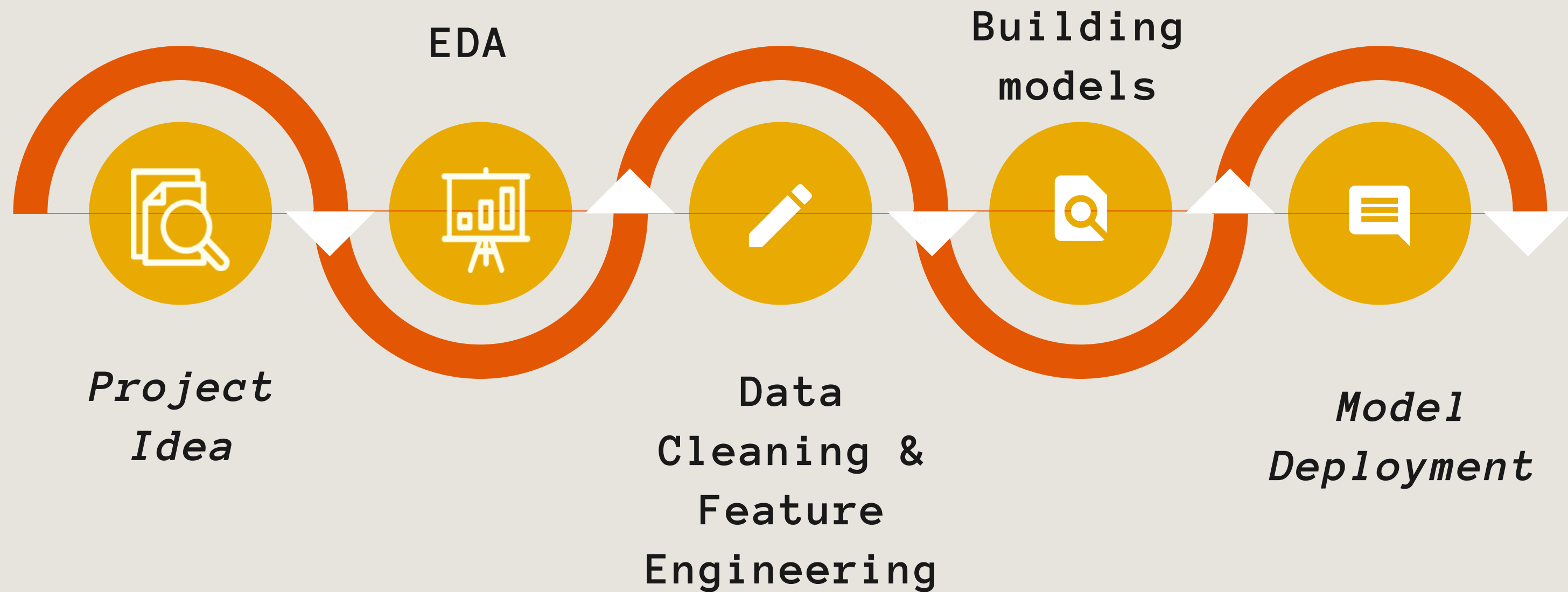
# 1. Introduction

# Introduction

- Will the employee work for the company or look for a new job ?
- It helps to reduce the cost and time as well as the quality of training or planning the courses

## **2. Workflow**

# WORKFLOW



# 3. Data & Design

# Dataset



## Dataset

- From Kaggle



## Size

- 19158 record
- 14 Features

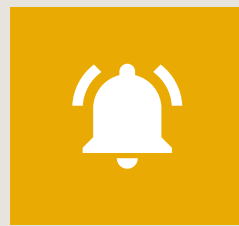


## Target

- Looking for a new job
- Not Looking for a new job



# DATA CLEANING & EDA



**Check null.**

**Deal with outlier.**

**Remove  
Unnecessary  
columns.**

**Drop  
duplicates.**

**Converting  
categorical  
values into  
numeric  
values**

# Data Preparation



## Feature Selection

- Drop 'enrollee\_id' and 'city' columns



## Feature Engineering

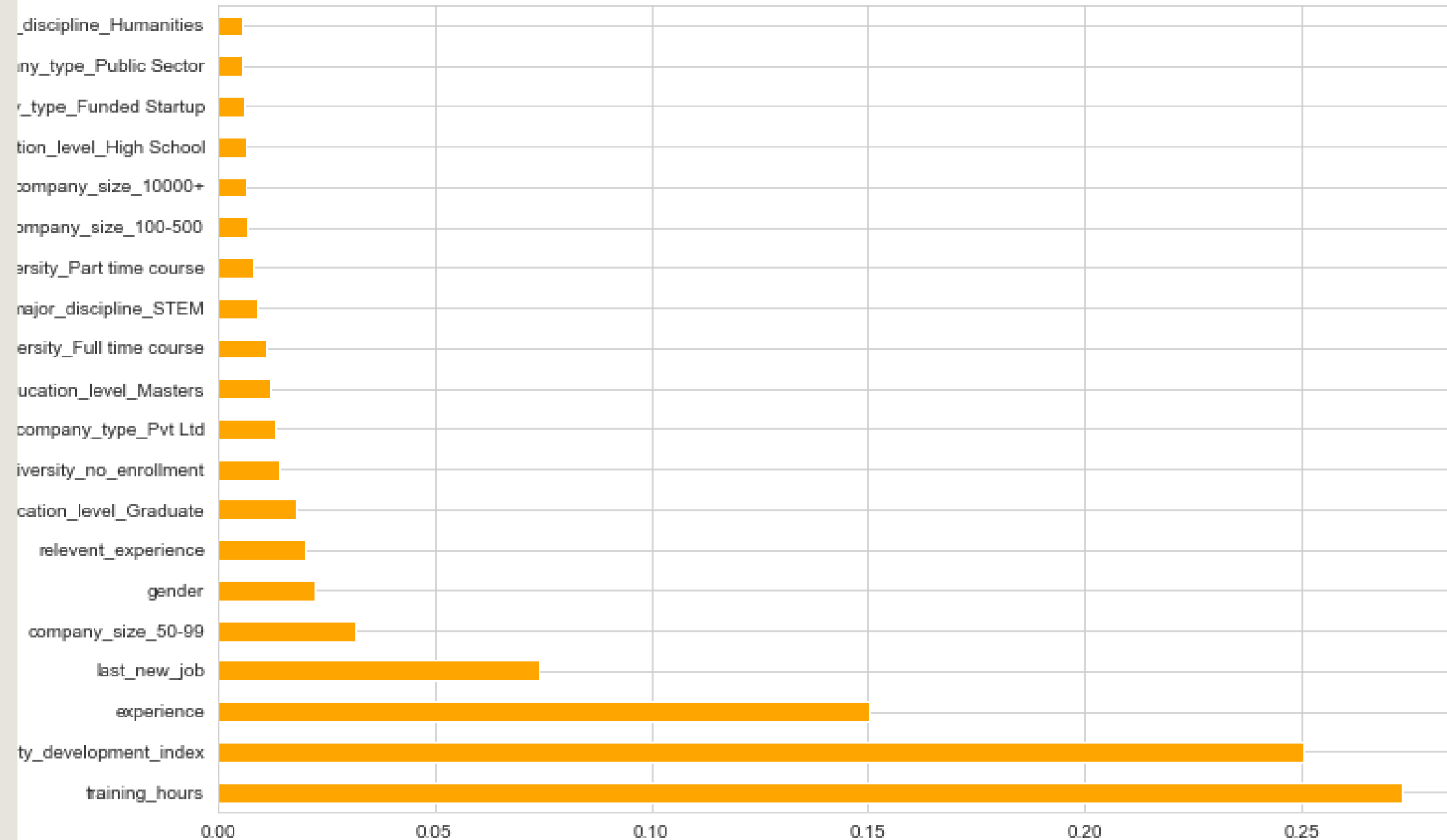
- label encoding
- get\_dummies (Encoding the columns into categorical values.)



## Imbalanced dataset

- SMOTE was use for handling the imbalanced

# Feature importance

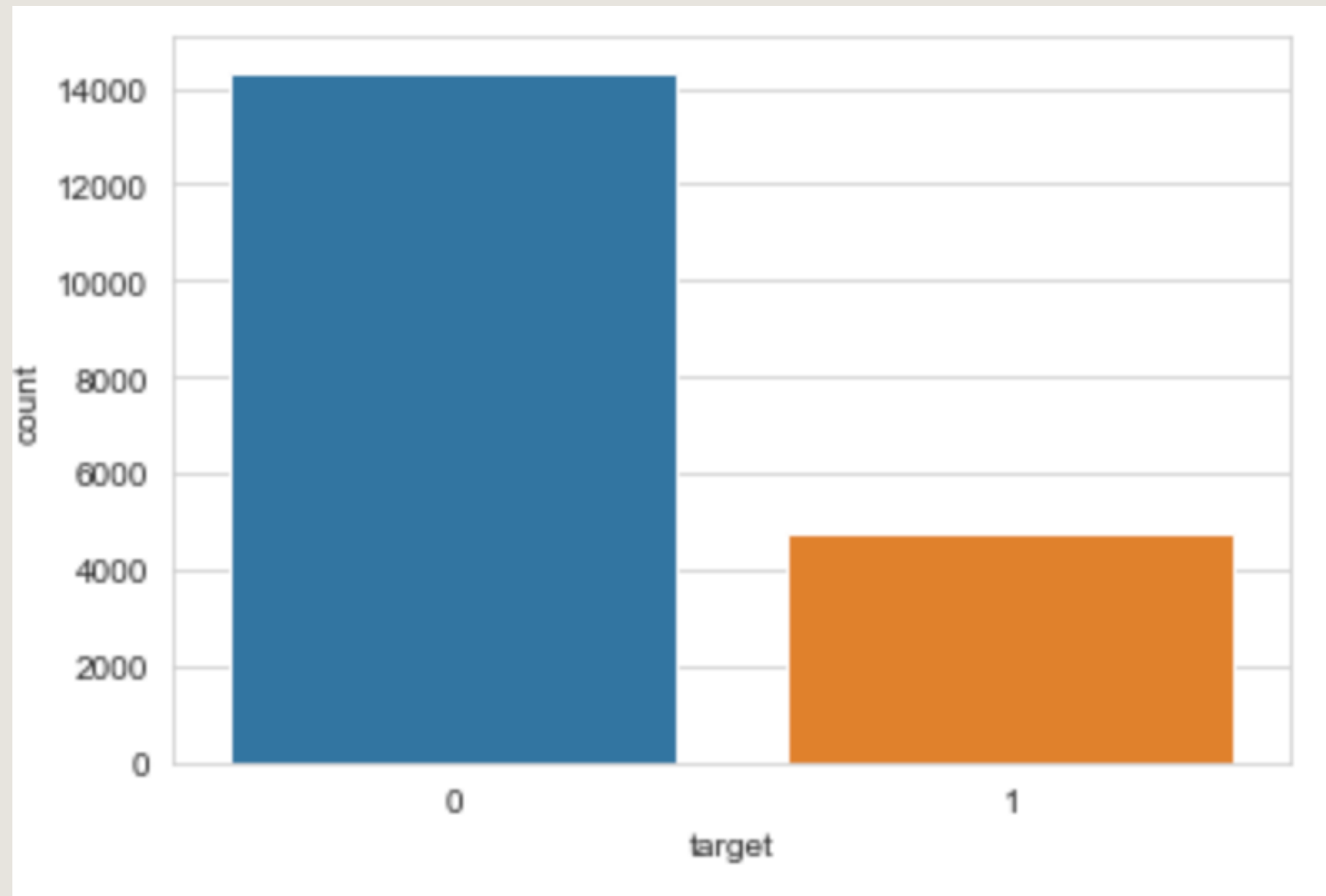


**According to bar chart,  
these featurers:**

- 1- training\_hours**
- 2- city\_development\_index**
- 3- experience**
- 4- last\_new\_job**
- 5-company\_size\_50-99**
- 6- gender**

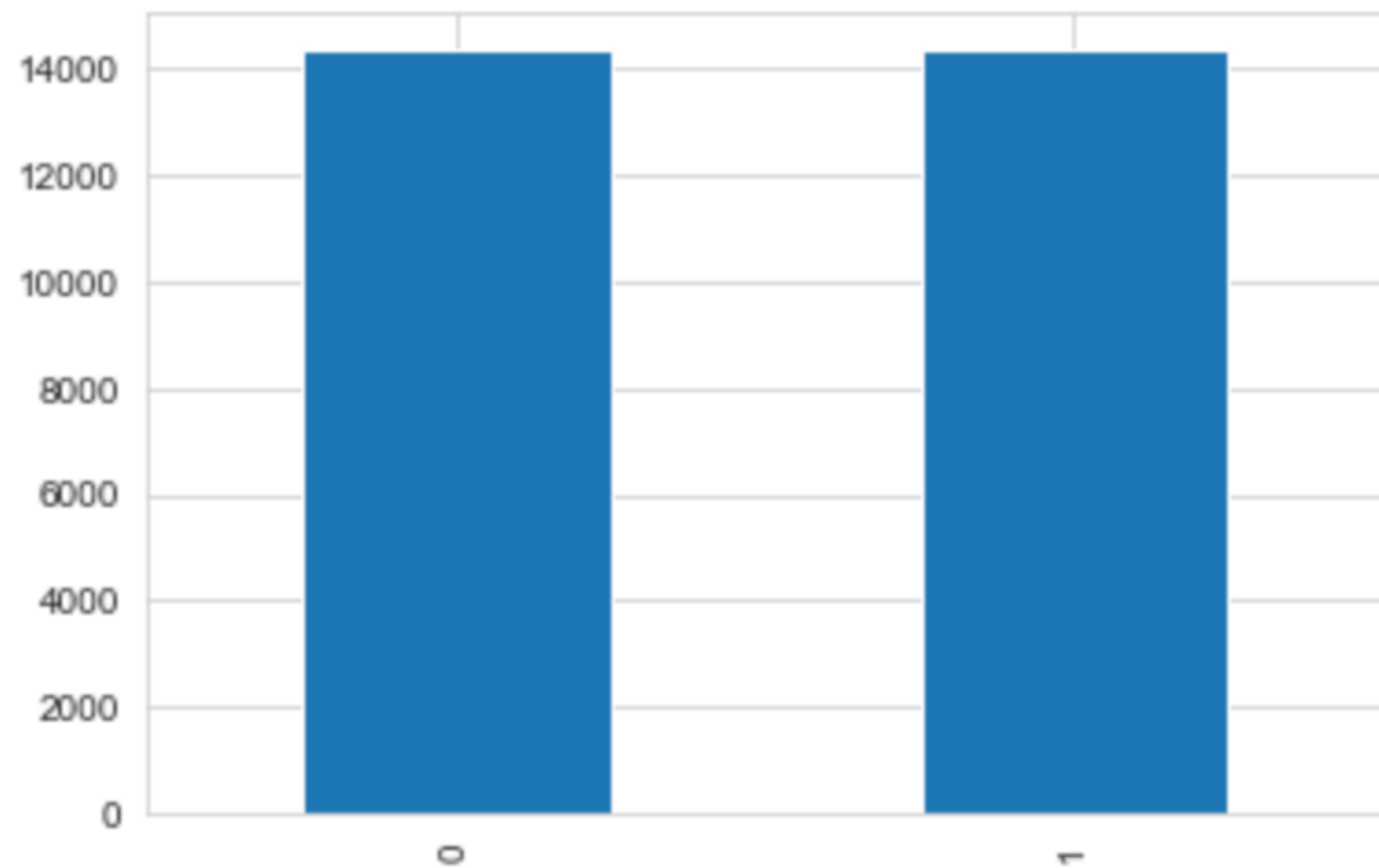
**Are the most important  
features .**

# Class Imbalance



*Target*  
**0: 14337**  
**1: 4761**

# Solving Class Imbalance



**SMOTE**

**0**: 14337

**1**: 14337



**ADAYSN**

**0**: 14337

**1**: 14337



**Random over  
sampler**

**0**: 14337

**1**: 14337

# 4.Models

# Classification Models

## F1 Score

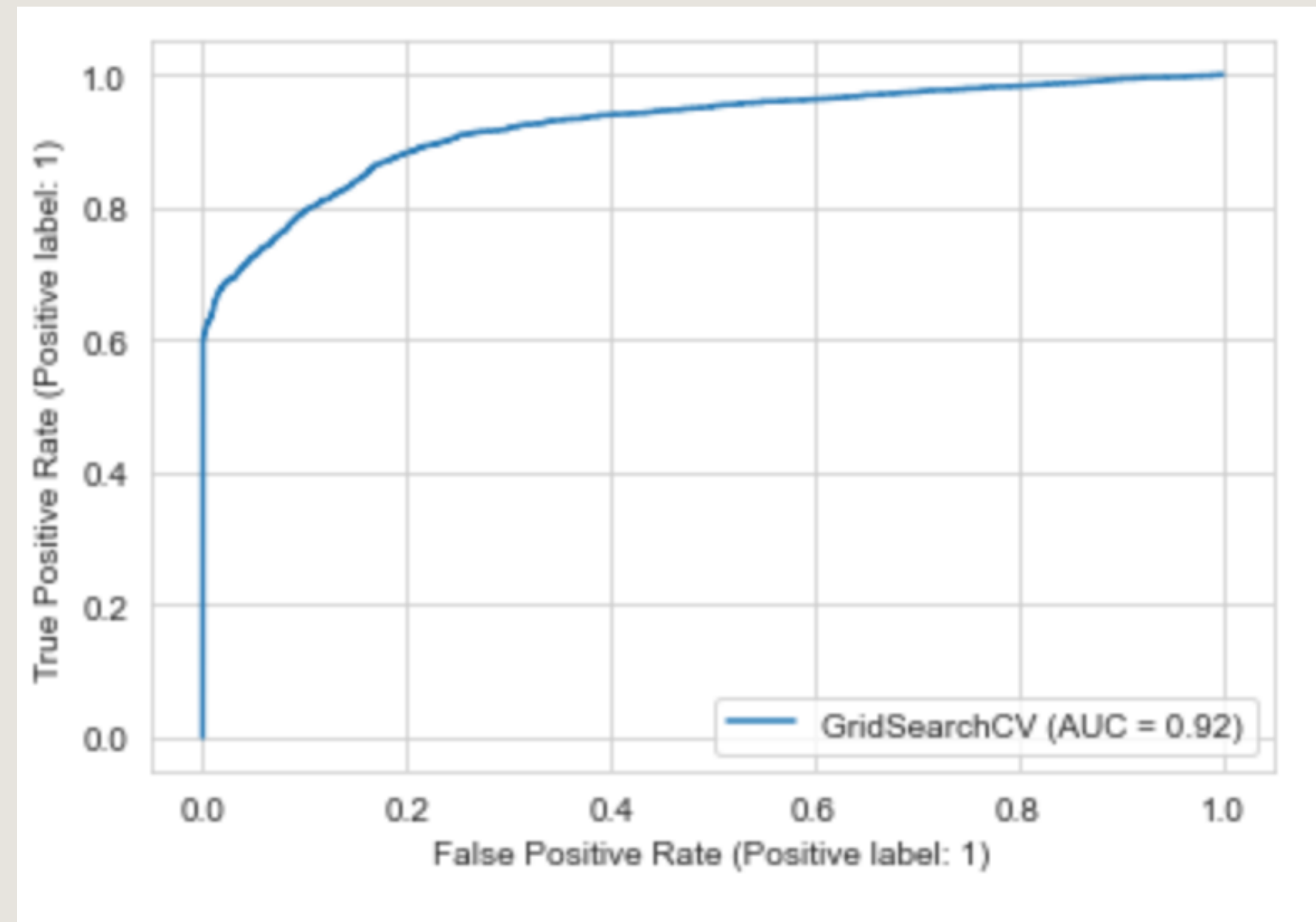
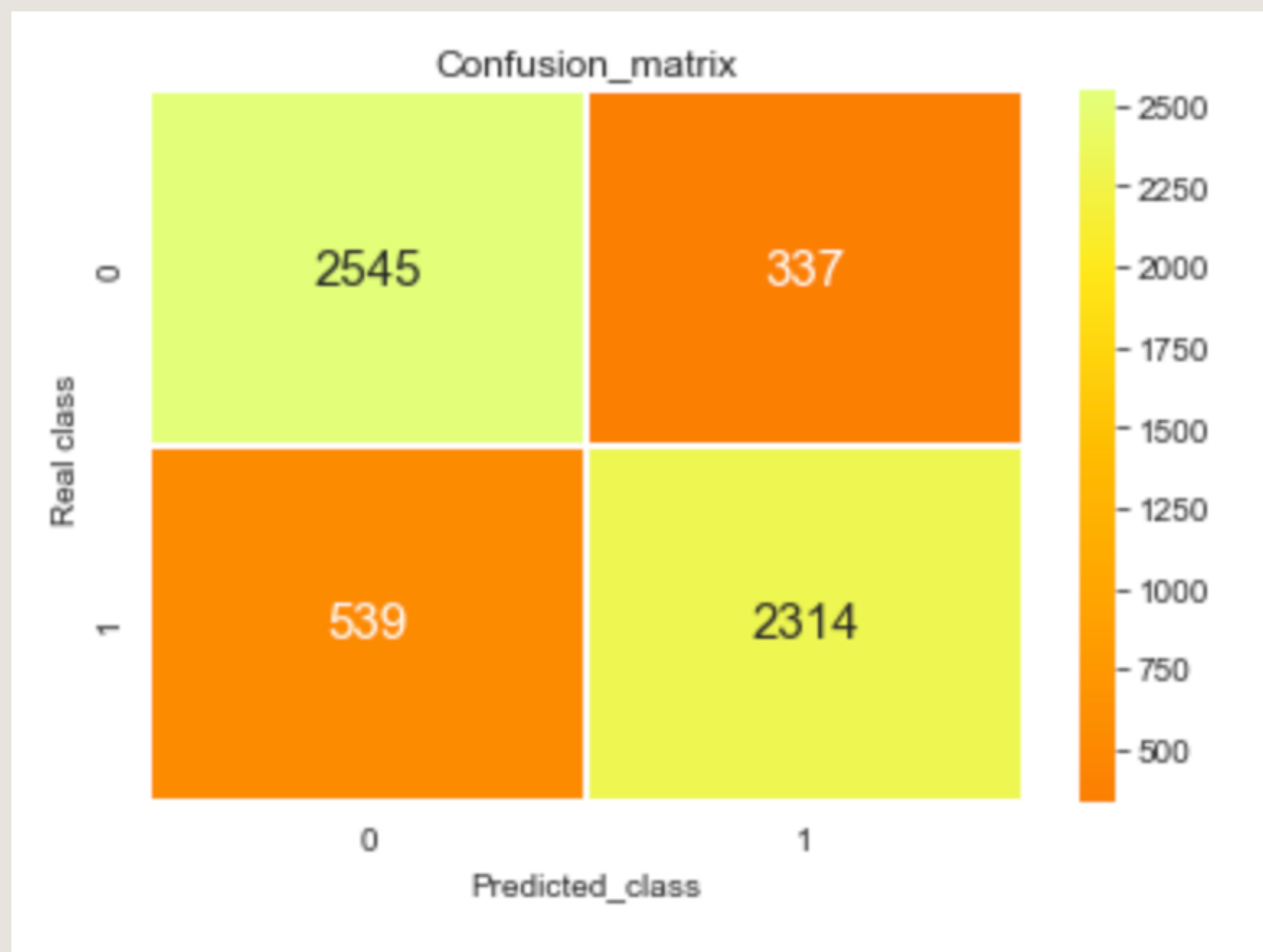
	Train	Validation	Test
Baseline Model	0.77	0.77	-
Logistic Regression	0.77	0.77	0.77
KNN	0.99	0.75	0.76
Decision Tree	0.81	0.80	0.80
Random Frost	0.79	0.81	0.75
XGBoost	0.95	0.91	0.92
SVC	0.58	0.58	0.58
GradientBoosting Classifier	0.84	0.83	0.83
AdaBoost Classifier	0.81	0.80	0.81
MLP Classifier	0.83	0.78	0.78

# The Best F1 Score Model

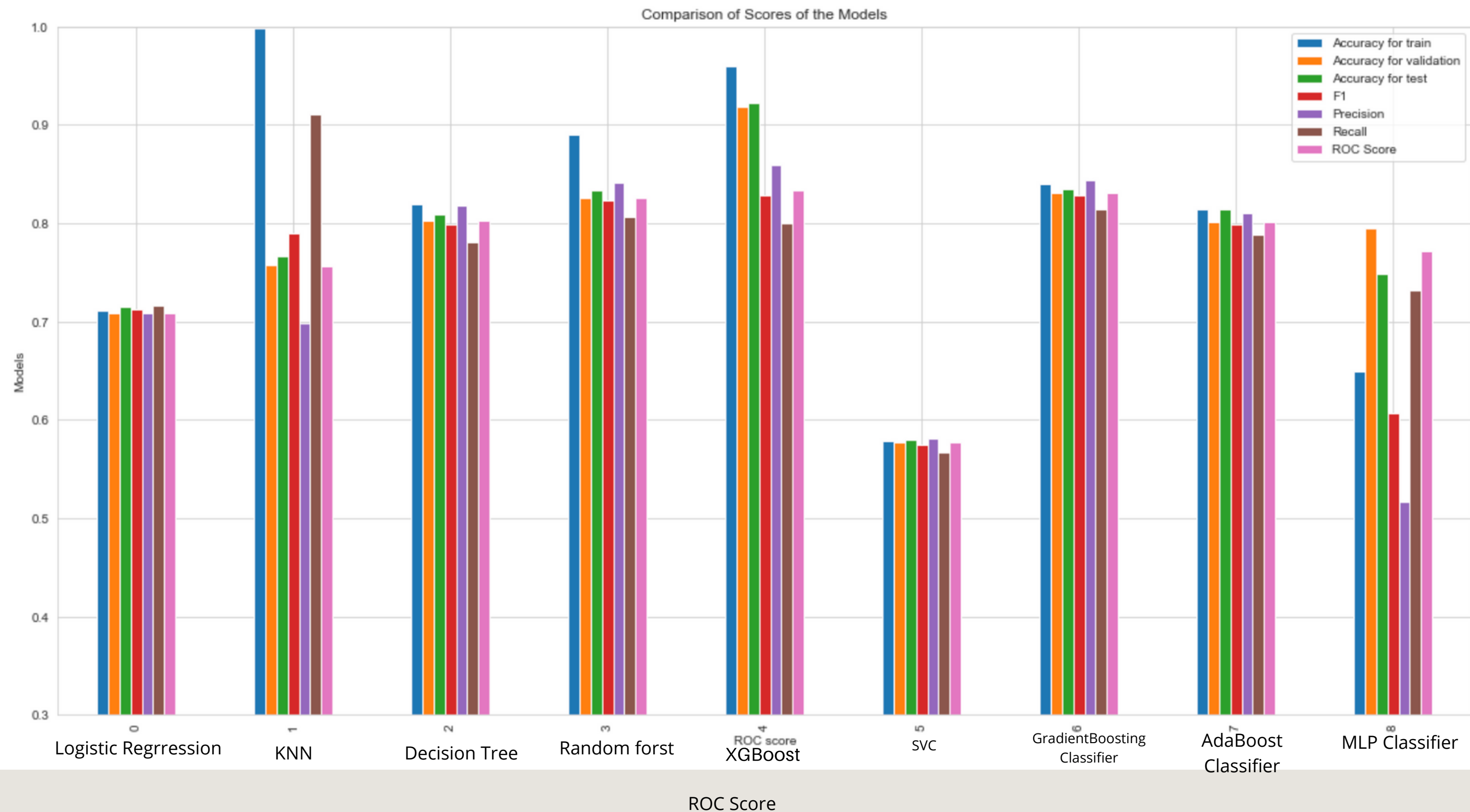
	Train	Validation	Test
Baseline Model	0.77	0.77	-
Logistic Regression	0.77	0.77	0.77
KNN	0.99	0.75	0.76
Decision Tree	0.81	0.80	0.80
Random Frost	0.79	0.81	0.75
XGBoost	0.95	0.91	0.92
SVC	0.58	0.58	0.58
GradientBoosting Classifier	0.84	0.83	0.83
AdaBoost Classifier	0.81	0.80	0.81
MLP Classifier	0.83	0.78	0.78



# Confusion Matrix and ROC (XGBoost)



# Comparison Between Models



# 5. Model Deployment

# Model Deployment

## Predict Data Scientist Job Change

Training hours:

50

Experience

3

Last new job

1

Gender(0 Male , 1 Female)

0

Predict

Employee will be looking for a new job

# 6.Tools

# Technologies and Libraries



Numpy , Pandas

**Data Cleaning &  
Manipulation**



Sklearn

**Model Building**



Matplotlib , Seaborn

**Visualization**



Flask

**Web Deployment**

# 7.Conclusion

# Conclusion

XGBoost provided the best prediction  
with accuracy score 0.92

## *Future Work;*

- *Optimizing the model*
- *Explore additional features*



**Thank you**